

A Performance Estimator for Quantum Annealers: Gauge selection and Parameter Setting.

Alejandro Perdomo-Ortiz,^{1,2,*} Joseph Fluegemann,^{1,3} Rupak Biswas,⁴ and Vadim N. Smelyanskiy¹

¹Quantum Artificial Intelligence Lab., NASA Ames Research Center, Moffett Field, CA 94035, USA

²University of California Santa Cruz at NASA Ames Research Center, Moffett Field, CA 94035, USA

³San Jose State Research Foundation at NASA Ames Research Center, Moffett Field, CA 94035, USA

⁴Exploration Technology Directorate, NASA Ames Research Center, Moffett Field, CA 94035

(Dated: March 4, 2015)

With the advent of large-scale quantum annealing devices, several challenges have emerged. For example, it has been shown that the performance of a device can be significantly affected by several degrees of freedom when programming the device; a common example being gauge selection. To date, no experimentally-tested strategy exists to select the best programming specifications. We developed a score function that can be calculated from a number of readouts much smaller than the number of readouts required to find the desired solution. We show how this performance estimator can be used to guide, for example, the selection of the optimal gauges out of a pool of random gauge candidates and how to select the values of parameters for which we have no *a priori* knowledge of the optimal value. For the latter, we illustrate the concept by applying the score function to set the strength of the parameter intended to enforce the embedding of the logical graph into the hardware architecture, a challenge frequently encountered in the implementation of real-world problem instances. Since the harder the problem instances, the more useful the strategies proposed in this work are, we expect the programming strategies proposed to significantly reduce the time of future benchmark studies and in help finding the solution of hard-to-solve real-world applications implemented in the next generation of quantum annealing devices.

I. INTRODUCTION

The fabrication of scalable hardware architectures for quantum annealers [1, 2] to solve discrete optimization problems has sparked interest in quantum annealing algorithms [3, 4]. Current research studies focus on both fundamental and practical important questions, including the implementation of real-world applications [5–9], defining criteria for detecting quantum speedup and the computational role of quantum tunneling [10, 11], proposals for error-suppression schemes [12], benchmark studies comparing classical and quantum annealing [13–19], and using spin-glass perspectives into the hardness of computational problems studied [20, 21].

The next generation of quantum annealers will likely allow for the exploration of harder and more interesting problems instances. Even in the case of a quantum processor with only ~ 500 qubits, one could already see the appearance of some hard to solve instances, for which the optimal solution was not found out of a few thousand annealing cycles [10]. It is precisely for these hard instances that the methods developed in this paper are the most useful, since they allow us to extract the best programming settings enhancing the probability of finding the ground state, therefore reducing significantly the time to solution for both future benchmark studies and for real-world applications.

The first step of solving a problem using a quantum annealer is to map the problem to the hardware

architecture. The quantum hardware employed consists of 64 units of a recently characterized eight-qubit unit cell [2, 22]. Post-fabrication characterization determined that only 509 qubits out of the 512 qubit array can be reliably used for computation (Fig. 4 in Appendix A). The array of coupled superconducting flux qubits is, effectively, an artificial Ising spin system with programmable spin-spin couplings and transverse magnetic fields. It is designed to solve instances of the following (NP-hard [23]) classical optimization problem: Given a set of local longitudinal $\{h_i\}$ and an interaction matrix $\{J_{ij}\}$, find the assignment $\mathbf{s}^* = s_1^* s_2^* \cdots s_N^*$, that minimizes the objective function $E(\mathbf{s})$,

$$E_{\text{ising}}(\mathbf{s}) = \sum_{1 \leq i \leq N} h_i s_i + \sum_{1 \leq i < j \leq N} J_{ij} s_i s_j, \quad (1)$$

where, $|h_i| \leq 2$, $|J_{ij}| \leq 1$, and $s_i \in \{+1, -1\}$. Finding the optimal \mathbf{s}^* is equivalent to finding the ground state of the corresponding Ising Hamiltonian, $H_p = \sum_i h_i \sigma_i^z + \sum_{i < j} J_{ij} \sigma_i^z \sigma_j^z$ where σ_i^z are Pauli matrices acting on the i th spin.

Physical realizations of quantum annealing come with certain degrees of freedom affecting the performance of the quantum annealing device. Each realization of such degrees of freedom determine a unique *Hamiltonian specification* or realization of Eq. 1. Although there is some understanding of the several factors affecting the performance of quantum annealing devices [24], there is a need for concrete scalable strategies coping with the analog control error (ACE) intrinsic to physical hardware implementations. For example, to the best of our knowledge there is no known “rule-of-thumb” in the selection of such parameters, thus motivating our study. (we rule out the

* Corresponding author’s e-mail: alejandro.perdomoortiz@nasa.gov

only one wide spread in the community, in Appendix B.)

In the absence of noise or any miscalibration, the performance of the quantum device should be the same under any gauge realization [14]. Previous studies show that the current generation of D-Wave devices with hundreds of qubits is very sensitive to this selection [7, 14]. Some other degrees of freedom correspond to parameters we do not yet know *a priori* how to set. This is the case for penalty strength in the construction of quadratic unconstrained binary optimization (QUBO) Hamiltonians [7, 25] or penalties associated with the strength of the set of qubits defining a logical qubit in the QUBO graph to hardware graph procedure [26].

In Sec. II we present a strategy for tuning and optimizing a quantum annealing algorithm, finding the best parameters out of a pool of candidates and selecting the Hamiltonian specifications with the best performance. It is in this section where most of the new results are presented. For accessibility to the readers, we divided this section into two main threads. Readers interested only in gauge selection (such as those researchers interested in benchmark studies, for example, on random spin-glass instances) can find the procedures needed in Sec. IIA - IIC. For readers interested in more general real-world application instances, where parameters for embedding procedures and other penalties need to be set, we devote Sec. IID to discussing the adjustments to the technique to deal with these additional challenges. In Sec. III, we delineate some future directions and possible further applications of the present work.

II. TUNING A QUANTUM ANNEALING ALGORITHM

As previously discussed, there are many degrees of freedom at the time of programming a quantum annealing device to solve a specific problem instance. Each realization of such degrees of freedom determine what we call a *Hamiltonian specification* for the quantum annealing cycles. For the purpose of generality, we leave the discussion at a very high-level form and in the following sections we will present application examples in different common practical scenarios, e.g., gauge selection and setting the strength of couplings among physical qubits representing a qubit from the original logical graph, also known as the embedding parameter setting problem [27]. In this general framework presented here, we only need to keep in mind that the performance of the device is determined by the programming degrees of freedom through the different Hamiltonian specifications. The main question we discuss next: How do we select the Hamiltonian realization that yields the best performance of the device? It is the focus of this section to answer this question with a procedure requiring a minimum overhead, as described next.

A. Performance Estimator: The Elite Mean, Π_{elite}

Assume that for each Hamiltonian specification you can easily request a total number of readouts N_{reads} from the quantum annealer. In some cases, these N_{reads} must be obtained in batches due to programming limitations. For example, in the current D-Wave Two processor hosted at NASA Ames, programming the device with an annealing time per cycle of $t_a = 100\mu s$ allows for a maximum number of readouts of 10,000. If the device is operated at $t_a = 20\mu s$ this maximum number is 50,000¹. Therefore, while at $t_a = 20\mu s$ a goal of $N_{\text{reads}} = 50,000$ can be obtained in one shot, at $t_a = 100\mu s$ we need to request 5 repetitions of 10,000 each. Let's denote the number of repetitions needed by n_{reps} and therefore the number of readouts in each repetition is $n_{\text{reads}} = N_{\text{reads}}/n_{\text{reps}}$.

For each readout, $\mathbf{s}^{(i)}$ there is a corresponding $E_{\text{ising}}(\mathbf{s}^{(i)})$ (Eq. 1). Let's define by $\tilde{\mathbf{E}}_{\text{ising}}$ as the array containing the n_{reads} sorted energies, i.e., $\tilde{\mathbf{E}}_{\text{ising}} = \{e_1, e_2, \dots, e_{n_{\text{reads}}}\}$ such that $e_i \leq e_j$ for all $j > i$. Define $\pi_{\text{elite}}^{\epsilon\%}$ as the negative of the mean value of the lowest ϵ percent of the energies in $\tilde{\mathbf{E}}_{\text{ising}}$. Since the array $\tilde{\mathbf{E}}_{\text{ising}}$ contains n_{reads} sorted energies from lowest to highest, then this expectation value is equivalent to calculating the mean value using the first $n_{\text{elite}} = \lceil \epsilon * n_{\text{reads}}/100 \rceil$ values in $\tilde{\mathbf{E}}_{\text{ising}}$. Formally defined,

$$\pi_{\text{elite}}^{\epsilon\%}(n_{\text{reads}}) = - \sum_{i=1}^{n_{\text{elite}}} e_i \quad (2)$$

Since only a fixed percent of the lowest energy values are included in the calculation, we refer to this score function hereafter as the *elite mean*. This expression can be generalized to the case where several repetitions are used to collect the desirable total number of samples N_{reads} by defining

$$\Pi_{\text{elite}}^{\epsilon\%}(N_{\text{reads}}, n_{\text{reps}}) = \frac{1}{n_{\text{reps}}} \sum_{i=1}^{n_{\text{reps}}} \pi_{\text{elite}}^{\epsilon\%} \left(\frac{N_{\text{reads}}}{n_{\text{reps}}}, i \right). \quad (3)$$

The minus sign in the definition Eq. 2 gives $\pi_{\text{elite}}^{\epsilon\%}$ the interpretation of a score function or a performance estimator; the higher its value, the better the expected performance. Suppose one has several quantum annealers or several Hamiltonian specifications to choose from. If one is interested in assessing the performance of the device with a number of reads $N_{\text{reads}} \ll R_{.99}$ (where $R_{.99}$ is defined as the number of readouts needed to find the desired solution at least once with a 99% probability), we will show that $\Pi_{\text{elite}}^{\epsilon\%}$ serves as an effective score function

¹ In our machine, there is a maximum duty time per submission. This value is set to $10^6\mu s$, the reason why the maximum number of readouts at $t_a = 20\mu s$ is 50,000 while only 10,000 at $t_a = 100\mu s$

or performance estimator that can be used to rank and to select the best of available quantum annealing specifications to solve the problem at hand. The intuition for this score function follows from what is expected of a quantum annealing device: when given a problem to be solved, the quantum annealers (or the Hamiltonian specifications) that give the lower energy solutions are preferable, since a quantum annealer is designed to sample from the lowest energy configurations. Therefore, the quantum annealer specification with the lower *elite mean energy* (or higher elite-mean score $\Pi_{\text{elite}}^{\epsilon\%}$), will give better performance.

B. Performance Rank

In quantum annealing, the most natural gold standard for assessing performance is the probability of observing the ground state, since it translates into the probability of finding the optimal solution to the optimization problem studied. In more precise terms, let's define the success probability of our quantum annealing algorithms by $p_s = n_{\text{gs}}/N_{\text{total}}$, where n_{gs} corresponds to the number of observed ground states in the total number of requested readouts N_{total} . Given p_s , the number of repetitions needed to observe the optimal solution at least once with a 99% probability, $R_{.99}$ is given by [14],

$$R_{.99} = \left\lceil \frac{\ln(1 - 0.99)}{\ln(1 - p_s)} \right\rceil \quad (4)$$

The instances that will benefit the most from our selection approach are those hard-to-solve instances with a very low probability of obtaining the ground state, say with an $R_{.99}$ in the order of hundreds of millions or hundreds of millions like the example discussed in Sec. IID ($R_{.99} \gg N_{\text{reads}}$). The purpose of this work is to show a correlation of the performance estimator and the real performance of the machine. But how do we define or assess real performance when the number of ground states is not reasonably attainable for all the Hamiltonian configurations we explored? Take for example the instance in Sec. IID. The default setting of the device does not provide even a single ground state solution after $N_{\text{total}} = 50 \times 10^6$ readouts! To calculate p_s for all gauges we would need to run for all 100 gauges being considered at least $N_{\text{total}} > 50 \times 10^6$, which is beyond the scope of this work. In this work we explore two definitions of performance that allow us to rank different Hamiltonian configurations even in the case where the ground state is not obtained after a significantly large N_{total} . The first and natural criteria is a *greedy-like performance rank*. This method gives a lower (better) rank to a Hamiltonian specification with a lower energy. In the common case of ties, they are broken by looking at the frequency (number of occurrences) of their lowest energy state. In case these are the same, the next lowest energy is compared and if they are still the same, one compares

their frequencies. The process continues until ties are broken, providing winners that are accordingly ranked lower. This method allows us to assign a unique ranking to any Hamiltonian specification whether or not we measured any ground states. Notice that in the particular case where the ground state is obtained for all the Hamiltonian configurations explored, the performance rank will still assign lower ranks to Hamiltonian specifications with larger values of p_s , as desired, while breaking any ties that exist.

C. Gauge selection: Case study with a random spin-glass instance

Benchmark studies assessing the presence or absence of speed-up of quantum annealers compared to classical processors resort to gauge selection as a way of obtaining reliable averaged results of the performance of the device [7, 10, 14, 24]. Although it is known that gauge specification can significantly enhance the performance of the device, previous studies are limited to the scaling of the typical gauge since there is no *a priori* way to determine the optimal gauge. Gauge specification is a particular example of Hamiltonian specification discussed above. We present here how our performance estimator $\Pi_{\text{elite}}^{\epsilon\%}$ can be used to select the optimal gauges. To illustrate the procedure, we used a hard-instance out of a pool of random-spin glass instances similar to the ones reported elsewhere [10]. This instance was provided by Sergio Boixo, who assessed that this particular instance had a simulated-annealing (SA) runtime of the order of hundred times longer than the median instance, from a pool of hundreds of thousands of instances within the same family (instances with random couplings $J_{ij} \in \{+1, -1\}$, with 509 qubits as shown in Fig. 4). As an abbreviation, we refer hereafter to this specific random-spin glass example as the RS instance. For QA this instance was shown to also be particularly hard after not obtaining any ground states after trying 16 gauges, with 10,000 readouts each, at 20 μs . Fig. 1(b) corroborates this assessment; the median gauge over a set of 100 random gauges has a $p_s = 1/(2 \times 10^6) = 5 \times 10^{-7}$, resulting in an expected number of repetitions to solution of $\bar{R}_{.99} \sim 9.2 \times 10^6$ annealing cycles. Since $\bar{R}_{.99}$ is about two hundred times greater than $N_{\text{reads}} = 50,000$, applying our performance estimator to select the optimal gauge before engaging in lengthier runs is expected to significantly reduce the computational time.

Fig. 1(a) shows there is a strong correlation between the rank obtained with $\Pi_{\text{elite}}^{2\%}$ with $N_{\text{reads}} = 50,000$ and $n_{\text{reps}} = 1$, compared to the greedy performance rank described above. The number of total readouts used to estimate the performance rank is 2 million per gauge. The error bars correspond to the rank provided by the first and third quartile out of 40 different experiments each with $N_{\text{reads}} = 50,000$ for each of the 100 random gauges. The middle point corresponds to the median of the set of

experiments. Fig. 1(b) shows the same data set but with the raw values for $\Pi_{\text{elite}}^{2\%}$ and also serves the purpose of showing the count in the number of ground states, illustrating that gauge selection can have a significant impact in the device performance, an increase as much as one to two orders of magnitude (see also Fig. 1(d) and Fig. 2(b) reflecting how the gauge selection can influence $n_{\text{gs.}}$)

As expected, any performance estimator would be a noisy metric and not expected to have a 100% correlation with the real performance from an extensive number of readouts $N_{\text{total}} \sim R_{.99}$. The RS section of Table I (upper half) addresses this issue. Suppose one utilizes the following strategy. One decides to run 100 gauges with a fixed N_{reads} for each gauge. From this starting data set, and while processing the readouts in the search of an optimal solution, one can easily calculate $\Pi_{\text{elite}}^{\epsilon\%}$. Since $N_{\text{reads}} \ll R_{.99}$, it is unlikely that this initial batch of calculations would contain the optimal solution. The refinement we propose here consists of using the information of the $\Pi_{\text{elite}}^{\epsilon\%}$ calculated *on-the-fly* to select, for example, the top 5 gauges (gauges with the highest $\Pi_{\text{elite}}^{\epsilon\%}$ score) out of the 100 random gauges. Since the selected gauges are expected to have a better performance than the typical or average gauge, only the selected ones are used to continue with the remaining runs $N_{\text{total}} \sim R_{.99}$ until the desired solution is found. Given this strategy, Table I answer the question: what is the probability that the absolute top gauge (that is, ranked number 1 according to the performance rank in Sec. IIB) is contained in this set of predicted top 5 gauges? What is the probability of one finding any of the top 2 gauges in the set of predicted top 5 gauges? etc, etc. Notice that at the level of $N_{\text{reads}} = 50,000$ which is ~ 200 times less than $R_{.99}$, one obtains a reasonably high $\sim 75\%$ probability of obtaining the top 1 gauge in the set top 5 gauges predicted by $\Pi_{\text{elite}}^{2\%}$. Table I also addresses the question of the existence of an optimal value of ϵ for the performance estimator. In all the examples considered here it seems to be the case that a value of $\epsilon\% = 1\%$ or 2% is optimal, a non-trivial result, since one might think incorrectly that the greater the number of low energy states included in the calculation of $\Pi_{\text{elite}}^{2\%}$ the better. This table also shows the expected increase in the probability of choosing the top gauges as N_{reads} becomes larger. Note also the inclusion in the table of a “Greedy” column for each case. This new metric was included because the use of the greedy method for the *performance rank* described in Section B begs the question of why one could not use an even simpler performance estimator consisting of the same *greedy approach* applied here to the case of a small number N_{reads} instead of N_{total} . However, the table clearly shows that $\Pi_{\text{elite}}^{2\%}$ is consistently always as good as, if not much better than, the greedy performance estimator. This is not surprising since as expected the $\Pi_{\text{elite}}^{2\%}$ should be a more robust metric than the greedy approach and to be less sensitive to rare-event occurrences. A clear advantage of the $\Pi_{\text{elite}}^{2\%}$ is that it provides a score function that can be used for other purposes as will be shown elsewhere [28].

The greedy approach as a score function is expected to be much flatter, due to this ranking relying heavily on the breaking of ties whenever there are only a few low energy states that many gauges reach).

D. Iterative strategy for embedding parameter setting and gauge selection

When solving real-world applications there are several additional subtleties to take into consideration. While in the case of the random spin-glass instances there is only one objective function appearing in the quantum annealing implementation, E_{ising} in Eq. 1, in instances derived from real-world applications E_{ising} is obtained from another cost energy function E_{QUBO} , a quadratic binary expression containing the logical qubits, $\{\vec{s}_{\text{logical}}\}$ before these are embedded into the hardware qubits, $\{\vec{s}_{\text{hardware}}\}$ appearing in E_{ising} [7]. From a practical application perspective, we are interested in the possibility of using our performance estimator to select the best Hamiltonian specification with the smallest $R_{.99}$, therefore reducing the computational time.

Once the embedding problem is solved [26, 27, 29] and one has a mapping of each logical qubit $s_{i,l} \in \{\vec{s}_{\text{logical}}\}$ into a subset of $\{\vec{s}_{\text{hardware}}\}$, the first decision to be made is to select the strength of the coupling J_E needed to keep the hardware spins representing a logical spin in alignment with each other.² For each value of J_E , and after requesting N_{reads} , we can calculate the fractions of the N_{reads} that do not violate any of these embedding constraints, i.e., with all the physical spins representing logical spins being properly aligned. These solutions are said to satisfy the *strict embedding* (SE) requirement. The right y -axis in Fig. 3(a) shows the fraction of solutions passing this requirement out of the total number of readouts N_{reads} , denoted as $f_{\text{SE}}(J_E)$.

Intuitively the magnitude of J_E cannot be too weak since it will not achieve the goal of keeping the spins properly aligned (same readout value for each variable representing the same logical qubit). Having a large value J_E will certainly help in increasing the probability of not having any misaligned spins but it cannot be too strong either, $J_E \gg \max(J_{ij})$, since after dividing everything by J_E to make all $-2 \leq h_i \leq 2$ and $-1 \leq J_E, J_{ij} \leq 1$, the original values h_i and J_{ij} will be well below the precision level and the performance will be significantly affected by noise. Therefore, a sweet spot with an optimal value of J_E is expected. From our experience, f_{SE} serves as a guide for selecting the region of interest, denoted here as $J_E^* < J_E < J_E^{**}$, with $f_{\text{SE}}(J_E^*) \approx 0.05$ and J_E^{**} corresponding to the onset of the plateau region with

² For simplicity we assume that all these penalties J_E enforcing the embedding are equal. Further fine tuning can be done by optimizing each parameter but this is beyond the scope of this work.

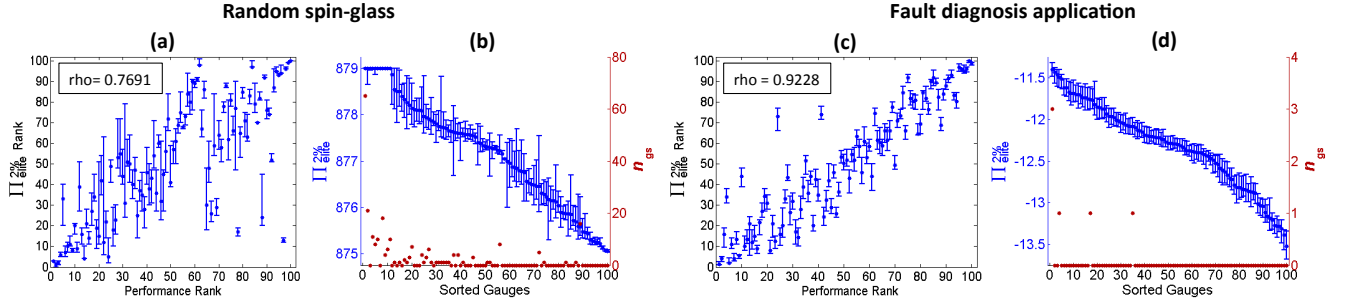


FIG. 1. **Correlation between Π_{elite} rank vs. performance rank:** (a-d) We refer to “one experiment” (a single experiment) as the ranking of the 100 gauges according to the values of $\Pi_{\text{elite}}^{2\%}$. In each of the 40 (100) experiments for the RS (50M-DMF) problem instance, each of the 100 gauges is scored and ranked according to $\Pi_{\text{elite}}^{2\%}$, obtained from $N_{\text{reads}} = 50,000$ per gauge. The median value of these experiments is shown as the middle point and the limits of the error bars correspond to the 25- and 75-percentile from these experiments at $t_a = 20\mu\text{s}$. The number of readouts ($N_{\text{reads}} = 50,000$) used to calculate $\Pi_{\text{elite}}^{2\%}$ rank in both examples is at least 100 times less than the number of repetitions required to find at least one solution with a probability of 99%, (i.e., $R_{.99} \gg N_{\text{reads}}$). (a,c) The Spearman coefficient (ρ) shows a moderately strong correlation of our estimator rank and the gold-standard rank used to define the performance. (b,d) Shown are the total number of ground states, n_{gs} , across all extensive runs from E_{ising} (EQUBO and using majority voting), with $N_{\text{total}} = 2$ millions ($N_{\text{total}} = 5$ million) for the RS (50M-DMF) problem instance. It is worth mentioning that the case of no-gauge with a performance rank of 77 out of 100 in panel (c) not even a single ground state was obtained after 50 millions readouts. Therefore, in this example a very bad gauge with a rank ~ 100 could take a significantly large computational time.

TABLE I. Fraction of experiments where the performance estimator $\Pi_{\text{elite}}^{\epsilon\%}(N_{\text{reads}})$ predicted correctly the top 1, either the top 1 or 2 (labeled Any Top 2), and any among the top 1, 2, or 3 (labeled Any Top 3), etc, within the set of 5 top-ranked gauges out of 100 random gauges, for the cases of the random-spin (RS) and the hardest fault diagnosis (50M-DMF) problem instances described in the main text. The number of experiments is 400, 200, 80, and 40 for $N_{\text{reads}} = 5,000, 10,000, 25,000$ and $50,000$, respectively. Note the non-trivial dependence on the fraction included in the elite mean, considering values of $\epsilon = 1, 2, 5$, and 10% , and showing an intermediate optimal value around 1% or 2% . As explained in the text, the *greedy* approach corresponds to the $(1/N_{\text{reads}})^*100$ -percentile, i.e., rank based on lowest energy obtained and breaking ties with the frequency of the lowest energy states). Annealing time per readout for all experiments was $20\mu\text{s}$.

RS	Top 1					Any Top 2					Any Top 3					Any Top 4					Any Top 5					
N_{reads}	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	
5k	0.50	0.52	0.53	0.53	0.41	0.64	0.73	0.73	0.66	0.52	0.88	0.94	0.95	0.93	0.81	0.89	0.91	0.91	0.90	0.86	0.97	0.97	0.97	0.97	0.94	
10k	0.53	0.57	0.57	0.57	0.43	0.67	0.70	0.73	0.69	0.59	0.90	0.92	0.92	0.92	0.84	0.93	0.94	0.94	0.92	0.84	0.99	0.99	0.99	0.99	0.96	
25k	0.58	0.64	0.64	0.66	0.45	0.70	0.74	0.75	0.75	0.60	0.83	0.88	0.88	0.95	0.85	0.9	0.91	0.9	0.95	0.85	0.96	0.98	0.98	1.00	0.98	
50k	0.73	0.75	0.75	0.68	0.50	0.88	0.88	0.88	0.75	0.60	0.95	0.95	0.95	0.95	0.93	1.00	0.95	0.95	0.95	0.93	1.00	1.00	1.00	1.00	1.00	
50M-DMF	Top 1					Any Top 2					Any Top 3					Any Top 4					Any Top 5					
	N_{reads}	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%	Greedy	1%	2%	5%	10%
	5k	0.49	0.58	0.59	0.57	0.54	0.68	0.78	0.78	0.73	0.68	0.73	0.83	0.83	0.78	0.74	0.73	0.84	0.83	0.79	0.75	0.86	0.93	0.93	0.87	0.82
	10k	0.54	0.64	0.64	0.62	0.6	0.72	0.8	0.81	0.78	0.74	0.77	0.86	0.87	0.83	0.79	0.77	0.86	0.87	0.83	0.79	0.89	0.95	0.95	0.9	0.85
	25k	0.62	0.72	0.74	0.72	0.72	0.85	0.85	0.85	0.82	0.8	0.90	0.91	0.9	0.86	0.85	0.90	0.91	0.9	0.86	0.85	0.98	0.97	0.98	0.91	0.91
	50k	0.71	0.84	0.82	0.75	0.73	0.92	0.92	0.88	0.85	0.84	0.93	0.95	0.91	0.9	0.87	0.93	0.95	0.91	0.9	0.87	0.99	0.99	0.97	0.95	0.91

$f_{\text{SE}} \approx f_{\text{SE}}^{\text{max}}$ in the plot f_{SE} vs. J_E . The value $f_{\text{SE}}^{\text{max}}$ can be easily obtained experimentally in one-shot by setting $J_E \gg 1$, and one can use that value to search for J_E^* .

For the purpose of our discussion we selected two instances from the fault diagnosis application published elsewhere [7]. The first instance, referred hereafter as 300K-DMF, was selected because despite its implementation with only 81 hardware qubits, it is unusually hard³ when compared with other benchmark studies [10], yet has a success probability just high enough to allow for a sizable number of ground states even in the worst Hamiltonian specification,⁴ within a reasonable number of readouts set to $N_{\text{reads}} = 300,000$ per gauge or J_E considered. As shown in Fig. 2, a finite number of ground states for every single gauge and every value of J_E allows

to rank each of these Hamiltonian specifications by their gold-standard performance rank and to compare with the rank predicted from our performance estimator. To satisfy the condition $\bar{R}_{.99} \gg N_{\text{reads}}$, the $\Pi_{\text{elite}}^{5\%}$ per gauges was calculated by using only $N_{\text{reads}} = 100$. Fig. 2, shows that even with so few readouts, there is a strong correlation between our Π_{elite} score function and the number of ground states observed after $N_{\text{total}} = 300,000$. Compared to the harder instances where $N_{\text{reads}} = 50,000$, here we used an $\epsilon\%$ of 5% instead of 2% , since the latter would amount to computing the elite mean with only the two lowest values out of $N_{\text{reads}} = 100$. This makes the estimator too noisy and flat, analogous to the “Greedy” performance estimator that only uses the lowest value to rank gauges. As shown in Fig. 2, calculating the elite mean over the five lowest energies already gives a good correlation with n_{gs} .

³ For the 300K-DMF, median success probability of $\bar{p}_s = 129/300,000 = 4.3/100,000$ out of set of 100 random gauges

⁴ For the 300K-DMF, the smallest $p_s = 25/300,000 = 8.3/100,000$

The second instance, referred hereafter as 50M-DMF instance, serves the purpose of showing how our performance estimator can be used in a practical situation,

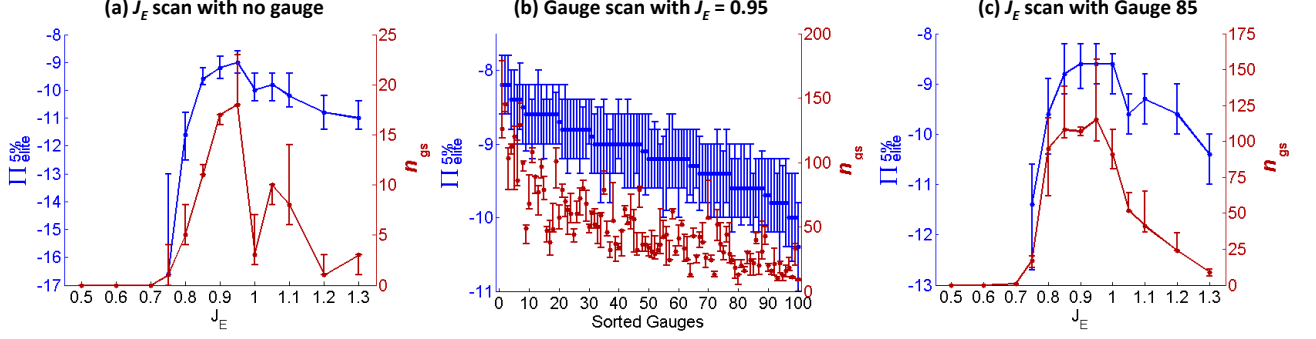


FIG. 2. The plots showing the strong correlation between our score function, $\Pi_{\text{elite}}^{5\%}$, and number of ground states, n_{gs} , for the 300K-DMF problem instance. The number of readouts, $N_{\text{reads}} = 100$, used to compute the score function is at least two orders of magnitudes less than the estimation of the typical number of readouts needed to observe the optimal solution at least once with a probability of 99% ($R_{.99} = 10,707$ for the median performance over 100 random gauges). The error bars in $\Pi_{\text{elite}}^{5\%}$ (blue) corresponds to the 25-percentile and the 75-percentile from 500 experimental realizations. The error bars in n_{gs} (red) corresponds to the lowest and highest of three realizations, each with $N_{\text{total}} = 100,000$, with the middle value as the median of the three experimental realizations. This shows that in these hard-instances, n_{gs} is still a noisy value, but still our performance estimator places us in the range of the top gauges with the largest n_{gs} . Each experimental realization consists of the estimation of $\Pi_{\text{elite}}^{5\%}(n_{\text{gs}})$ by using $N_{\text{reads}}(N_{\text{total}})$ per gauge. The iterative approach described here is the same one described in the text and in Fig. 3 to set the embedding parameter J_E and to select the optimal gauges.

for instances with probabilities much smaller. These are the instances we expect to surface in the next generation of quantum annealers. The instance 50M-DMF has the property of having a unique optimal solution, making it the most difficult to solve among the family of problem instances with six-faults to be diagnosed. More specifically, although the number of hardware qubits (96 qubits) required to implement this instance is not unusually large, this instance turned out to be extremely difficult for QA; not even a single-ground state was measured after $N_{\text{total}} = 50 \times 10^6$ annealing cycles, even after optimizing for the optimal J_E but under the default no-gauge! This instance was in large the motivation for defining a quick strategy to find the optimal Hamiltonian specifications (best J_E and best gauge) capable of finding the ground state).

Fig. 3 describes the suggested iterative approach used to optimize both the value of J_E and to select the optimal gauges in instances requiring a direct embedding approach. Starting with the no-gauge one can scan for the candidate values of J_E and select the value of J_E with the highest score $\Pi_{\text{elite}}^{2\%}$. In contrast with the case of the RS instance above, here we cannot use E_{ising} to calculate the score function since, for example, the lowest energies of E_{ising} will be different for every value of J_E (because of the energy renormalization to fit all programmable values h_i, J_{ij} , and J_E within the dynamical range $|h_i| \leq 2$ and $|J_{ij}| \leq 1$). To circumvent this issue we compute $\Pi_{\text{elite}}^{2\%}$ after error-correcting the N_{reads} solutions with majority voting [24] when going from $\{\tilde{s}_{\text{hardware}}\} \rightarrow \{\tilde{s}_{\text{logical}}\}$, and then sorting the states according to $E_{\text{QUBO}}(\{\tilde{s}_{\text{logical}}\})$ before selecting the 2% of the lowest energies used in the computation of $\Pi_{\text{elite}}^{2\%}$.

The next step in the procedure is to perform a *gauge scan* at the selected value of J_E from the J_E scans. Considering that we are dealing with instances with $R_{.99} \gg N_{\text{reads}}$, it is not a significant usage of compu-

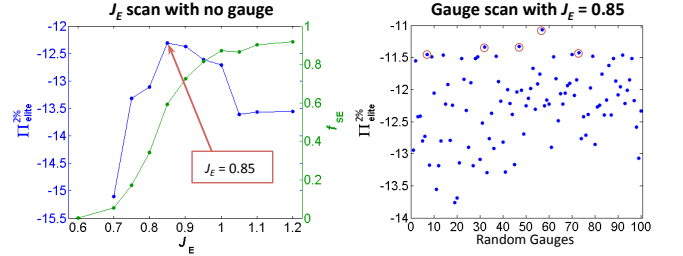


FIG. 3. **Iterative strategy for parameter setting and gauge selection for the 50M-DMF instance:** The corresponding data illustrates only one realization for the selection of J_E (left) and another for the selection of the top 5 gauges (circled) with the highest score (right). For a statistical analysis of the robustness of the method, see the second part of Table I. The left y-axis for the J_E plots the value of the $\Pi_{\text{elite}}^{2\%}$ when calculated using $N_{\text{reads}} = 50,000$ which is more than three-orders of magnitude less than the number of readouts to solution, $R_{.99}$, for this problem. The right y-axis for this J_E plot shows the percentage of solutions that passed strict embedding (solutions with no violations of the constraints imposed by J_E). f_{SE} serves as a guide for selecting the region with the optimal J_E , as explained in the main text.

tational resources to perform calculations with a number of gauges on the order of about 100. Notice that there is really no overhead while doing the gauge scans, since for every gauge considered, one needs to post-process all the solution readouts (e.g., with majority voting) while searching for the states with the optimal solutions anyways. Since the energies of every single solution needs to be calculated, the only overhead in calculating $\Pi_{\text{elite}}^{2\%}$ comes from a cheap sorting of these energies before calculating the elite mean. For NP-complete problems, we can always tell if we have found the desired answer. Also, for a large family of problems such as those NP-hard problems where the NP-complete version is still interesting, one can still stop the search if the desired solution is obtained (e.g., we can ask whether there exists a solution with an energy lower than a reference energy, with the latter being for example the best solution attainable with a state of the art classical solver). Therefore, trying ~ 100

gauges in the search of an optimal gauge is not an unfeasible idea. After calculating $\Pi_{\text{elite}}^{2\%}$ for the complete pool of gauge candidates, one can proceed to another set of J_E -scans by using the best gauge with the highest score. As shown in Fig. 2(c) and from our experience with other problem instances where the procedure was even applied at different annealing times, in most of the cases the second optimal J_E matched the same J_E from the first scan under the no-gauge setting. Even in the cases where J_E moved to a new value, the change was in the neighborhood of the first optimal value. As shown in Fig. 2(a) and (c), as long as one is near the optimal value of J_E the performance is not significantly affected. The gauge selection, Fig. 2(b), seems to have a much larger impact in the performance. Since the first J_E -scan does the job of taking us to the neighborhood of J_E optimality (out of the set of candidates considered), it is reasonable to conclude that a second J_E is not necessary and it is better to focus on the top gauges obtained from a gauge scan of 50-100 gauges. For easy instances a large gauge set would be unnecessary since the optimal solution will likely appear before one finishes going through the target number of 100 or so gauges.

As shown in Fig. 1, the performance estimator proposed here is a noisy metric. For example, there is no guarantee that the top gauge is the same one as the one predicted by $\Pi_{\text{elite}}^{2\%}$. Therefore, instead of selecting only the gauge predicted as top 1, it is advisable to select a handful of the predicted as top gauges as indicated in Fig 3(b). It is with this selected set that one performs the extensive runs $N_{\text{total}} \sim R_{.99}$, but where now $R_{.99}$ has been significantly reduced given that we are running with a set that includes the optimal gauge from the random set. Table I shows that selecting the predicted top five gauges has a high probability ($> 80\%$) of containing the top 1 gauge yielding the largest number of ground states. Predicting any of the top 2 gauges had a probability $\sim 90\%$. This is very remarkable considering that in this particular 50M-DMF problem running a low-performing gauge would lead to a significantly large time to solution. As mentioned above, the default no-gauge did not find the solution after 50 millions reads, therefore yielding a $R_{.99} > 230$ millions, while any of the top 3 gauges require $R_{.99} < 23$ millions, providing at least an order of magnitude improvement in this hard-to-solve instance for the QA processor.

In the case of real-world applications that use ancilla variables [7, 25] in the construction of their E_{QUBO} post-processing strategies are also possible. In these cases it is more efficient to process the solution, for example, evaluating the problem energy, E_{problem} with only the relevant variables defining the problem. More specifically, and without loss of generality, we can express the set of resulting logical qubits in the E_{QUBO} expression as $\{\vec{s}_{\text{logical}}\} = \{\vec{s}_{\text{problem}}\} \cup \{\vec{s}_{\text{ancilla}}\}$, where $\{\vec{s}_{\text{problem}}\}$ corresponds to the set of qubits or binary variables that define completely the problem description and that can be extracted to evaluate the energy of the problem, E_{problem} . In these

cases, and with the intention of increasing the chances of finding the optimal solution, it is more efficient to process the solution readouts with $E_{\text{problem}}(\{\vec{s}_{\text{problem}}\})$ and not with $E_{\text{QUBO}}(\{\vec{s}_{\text{logical}}\})$. This postprocessing strategy can only help in finding the optimal solution, since for every readout $E_{\text{problem}}(\{\vec{s}_{\text{problem}}\}) \leq E_{\text{QUBO}}(\{\vec{s}_{\text{logical}}\})$, therefore allowing for the possibility of finding optimal solutions in solutions that had been penalized by the ancilla constrains. Our preliminary results indicate that for these problem instances, it is advisable to look also at the top 5 gauges obtained from the greedy approach (same approach described in Table I but now using E_{problem} instead of E_{QUBO}) along with the top 5 from the Π_{elite} score function, also calculated with E_{problem} instead of E_{QUBO} . The strategy proposed here consists of taking as the “selected top gauges” the union set of these two sets of top 5 gauges, and perform with these gauges the extensive runs with $N_{\text{total}} \sim R_{.99}$.

III. CONCLUSIONS

We defined a score function intended to estimate the performance of quantum annealers whose applicability does not rely on obtaining ground states corresponding to the desired solution. We observed a strong correlation of our performance estimator with the performance of the device even in the case where the number of readouts used to calculate it was several orders of magnitude less than the number of readouts required to find the desired solutions. The score function is based on a tail conditional expectation value, corresponding to the *elite mean* over a small percent representing the readouts with the lowest energies. We showed it can be used to efficiently select the optimal gauges from a large pool of random gauges and in setting Hamiltonian parameters appearing in the implementation of real-world applications.

Although it has been previously shown that the decisions in programming quantum annealing devices can significantly impact the performance of the device [7, 14, 19], thus far comparison of performance of quantum annealers to algorithms on conventional classical processors was limited to average performance over the selection of parameters explored. This study opens the possibility of revisiting such scaling studies, now with the opportunity to select in advance the best configuration of the device. Having the possibility of selecting the specifications (best gauges or other optimal parameters) will be indispensable once we start solving instances intrinsically harder with the new generation of quantum annealers. The overhead incurred to apply the selection procedure presented here is constant and it does not scale with the size of the system. We showed that even in cases where $N_{\text{reads}} < R_{.99}/1000$, the method still works with a large probability of selecting the top gauges.

In the case of real-world applications, the iterative strategy proposed in Sec. II D requires essentially no overhead in calculating the performance estimator and used

to rank the random gauges. Since the data needs to be processed anyway (e.g., calculation of E_{QUBO} and majority voting while testing whether or not the desired solution has been found), the only overhead incurred is the time needed to sort the solution before calculating the elite mean. In the case of other parameter settings such as the one used in the embedding problem, our performance estimator provides a very efficient approach by pinning down the region where the device has its best performance.

Although our strategy allows us to select the best Hamiltonian specification in quantum annealers, we do not expect that it will be enough to change the complexity class seen in scaling studies [10, 14]. Certainly, it could easily provide a speed-up of an order of magnitude from the default methods, as seen in some of the examples presented here, and it might be to-date the only feasible way to obtain solutions to hard-to-solve computational problems (either random spin-glass benchmarks or real-world applications) in the next generations of quantum annealers.

ACKNOWLEDGEMENTS

This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IAA 145483. We want to acknowledge the support of NASA Advanced Exploration Systems program and NASA Ames Research Center. The authors thank Sergio Boixo for providing the 509-qubit RS problem instance, and Bryan O’Gorman, Eleanor Rieffel, and Davide Venturelli for helpful discussions.

AUTHOR CONTRIBUTIONS

A.P-O, J.F, R.B and V.N.S contributed to the ideas presented in the paper. A.P-O and J.F designed and ran the experiments and wrote the manuscript. All the authors revised the manuscript.

Appendix A: Chimera architecture

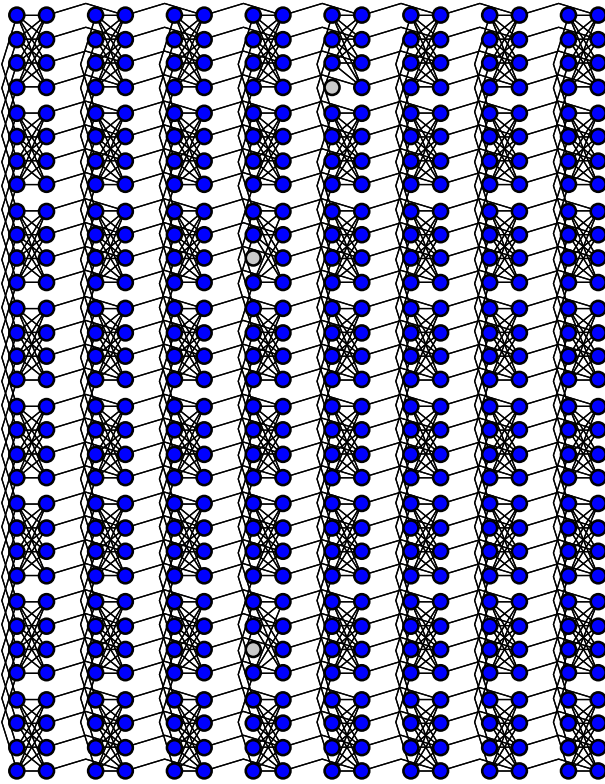


FIG. 4. **Device architecture and qubit connectivity D-wave Two at NASA Ames:** The array of superconducting quantum bits is arranged in 8×8 unit cells that consist of 8 quantum bits each. Within a unit cell, each of the 4 qubits in the left-hand partition (LHP) connects to all 4 qubits in the right-hand partition (RHP), and vice versa. A qubit in the LHP (RHP) also connects to the corresponding qubit in the LHP (RHP) of the units cells above and below (to the left and right of) it. Edges between qubits represent couplers with programmable coupling strengths. Blue qubits indicate the 509 usable qubits, while grey qubits indicate the three unavailable ones out of the 512 qubit array.

Appendix B: Evidence against a commonly used rule-of-thumb for gauge selection

In the case of gauge selection, a commonly used “rule-of-thumb” that had persisted in the community is that the gauge maximizing the number of antiferromagnetic couplings, $J_{ij} > 0$ is preferred. The physical motivation behind this “rule” is that the precision in the specification of a $J_{ij} > 0$ (antiferromagnetic coupling) is more robust than its negative (ferromagnetic) counterpart [30]. A more detailed analysis including 100 gauges for several problem applications considered (see Fig. 5) shows that such rule-of-thumb does not hold in any of the hard instances considered here. Notice there is no correlation between the number of positive couplers and the performance of the specified gauge. We did not see any correlation either in any other quantity similar to J_{ij} : parameters studied include the number of $J_E > 0$ (for the case of real-world applications with direct embedding), the number of $h_i > 0$ and the number of J_{ij} that are non- J_E . For all those cases, still no correlation was found.

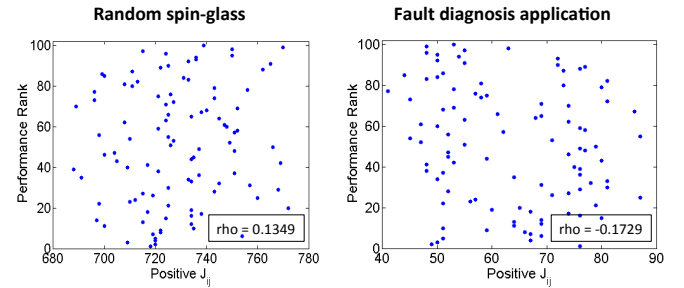


FIG. 5. The low value of spearman correlation coefficient shows that there is basically no correlation between the number of $J_{ij} > 0$ resulting from a specified gauge and the performance in the device, ruling out the common belief that the larger the number of $J_{ij} > 0$, the better. Shown here are examples from three different application domains.

- [1] Bunyk, P. *et al.* Architectural considerations in the design of a superconducting quantum annealing processor. *Applied Superconductivity, IEEE Transactions on* **24**, 1–10 (2014).
- [2] Johnson, M. W. *et al.* Quantum annealing with manufactured spins. *Nature* **473**, 194–198 (2011).
- [3] Kadowaki, T. & Nishimori, H. Quantum annealing in the transverse ising model. *Phys. Rev. E* **58**, 5355 (1998).
- [4] Farhi, E. *et al.* A quantum adiabatic evolution algorithm applied to random instances of an NP-Complete problem.

Science **292**, 472–475 (2001).

- [5] Perdomo-Ortiz, A., Dickson, N., Drew-Brook, M., Rose, G. & Aspuru-Guzik, A. Finding low-energy conformations of lattice protein models by quantum annealing. *Sci. Rep.* **2**, 571 (2012).
- [6] Gaitan, F. & Clark, L. Ramsey numbers and adiabatic quantum computing. *Phys. Rev. Lett.* **108**, 010501 (2012).
- [7] Perdomo-Ortiz, A., Fluegemann, J., Narasimhan, S., Biswas, R. & Smelyanskiy, V.N. A quantum anneal-

- ing approach for fault detection and diagnosis of graph-based systems. *Eur. Phys. J. Special Topics* **224**, 131–148 (2015).
- [8] Rieffel, E. G. *et al.* A case study in programming a quantum annealer for hard operational planning problems. *Quantum Information Processing* **14**, 1–36 (2015).
 - [9] O’Gorman, B., Babbush, R., Perdomo-Ortiz, A., Aspuru-Guzik, A. & Smelyanskiy, V. Bayesian network structure learning using quantum annealing. *Eur. Phys. J. Special Topics* **224**, 163–188 (2015).
 - [10] Rønnow, T. F. *et al.* Defining and detecting quantum speedup. *Science* **345**, 420–424 (2014).
 - [11] Boixo, S. *et al.* Computational role of multiqubit tunneling in a quantum annealer. *arXiv:1502.05754* (2015).
 - [12] Pudenz, K. L., Albash, T. & Lidar, D. A. Error-corrected quantum annealing with hundreds of qubits. *Nat Commun* **5** (2014).
 - [13] Boixo, S., Albash, T., Spedalieri, F. M., Chancellor, N. & Lidar, D. A. Experimental signature of programmable quantum annealing. *Nat Commun* **4** (2013).
 - [14] Boixo, S. *et al.* Evidence for quantum annealing with more than one hundred qubits. *Nature Physics* **10**, 218–224 (2014).
 - [15] Shin, S. W., Smith, G., Smolin, J. A. & Vazirani, U. Comment on ”distinguishing classical and quantum models for the d-wave device”. *arXiv:1404.6499v2* (2014).
 - [16] Albash, T., Rønnow, T. F., Troyer, M. & Lidar, D. A. Re-examining classical and quantum models for the d-wave one processor. *Eur. Phys. J. Special Topics* **224**, 111–129 (2015).
 - [17] Martin-Mayor, V. & Hen, I. Unraveling quantum annealers using classical hardness. *arXiv:1502.02494* (2015).
 - [18] Hen, I. *et al.* Probing for quantum speedup in spin glass problems with planted solutions. *arXiv:1502.01663* (2015).
 - [19] King, A. D. Performance of a quantum annealer on range-limited constraint satisfaction problems. *arXiv:1502.02098* (2015).
 - [20] Katzgraber, H. G., Hamze, F. & Andrist, R. S. Glassy chimeras could be blind to quantum speedup: Designing better benchmarks for quantum annealing machines. *Phys. Rev. X* **4**, 021008 (2014).
 - [21] Venturelli, D. *et al.* Quantum optimization of fully-connected spin glasses. *arXiv:1406.7553* (2014).
 - [22] Harris, R. *et al.* Experimental investigation of an eight-qubit unit cell in a superconducting optimization processor. *Phys. Rev. B* **82**, 024511 (2010).
 - [23] Barahona, F. On the computational complexity of ising spin glass models. *J. Phys. A: Math. Gen.* **15**, 3241–3253 (1982).
 - [24] King, A. D. & McGeoch, C. C. Algorithm engineering for a quantum annealing platform. *arXiv:1410.2628* (2014).
 - [25] Babbush, R., Perdomo-Ortiz, A., O’Gorman, B., Mcready, W. & Aspuru-Guzik, A. *Construction of Energy Functions for Lattice Heteropolymer Models: Efficient Encodings for Constraint Satisfaction Programming and Quantum Annealing*, 201–244 (John Wiley Sons, Inc., 2014).
 - [26] Cai, J., Mcready, B. & Roy, A. A practical heuristic for finding graph minors. *arXiv:1406.2741* (2014).
 - [27] Choi, V. Minor-embedding in adiabatic quantum computation: I. the parameter setting problem. *arXiv:0804.4884* (2008).
 - [28] Perdomo-Ortiz, A. *et al.* Determination of correctable persistent biases in quantum annealers. *In Preparation* (2015).
 - [29] Choi, V. Minor-embedding in adiabatic quantum computation: II. minor-universal graph design. *Quantum Information Processing* **10**, 343–353 (2011).
 - [30] *Personal communication with T. Lanting, K. Pudenz, and S. Adachi.* (2014).